

MTC Conceptual Data Model Strategy Planning

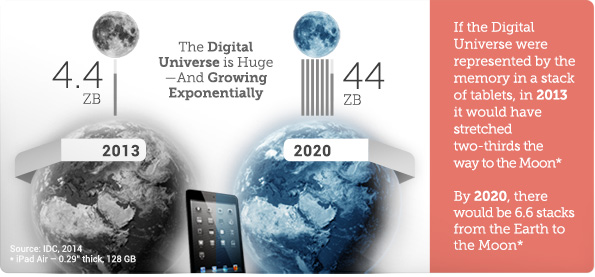
By

Victoria Giryavets and Michael Ziyambi

**Table of Content**

* Executive Summary
* Approach
* Conceptual Model Development Framework
* Results
* Conclusion

**Executive Summary**

Integrated Planning Section at MTC uses and maintains an extensive variety of data that is used to support growing number of projects and initiatives. Currently the data is spread between several databases, shared drives, box and personal drives in many different formats. “By 2020 IDC” [International Data Corporation] “predicts the amount of digital information will grow by a factor of 30, the number of files, by a factor of 60”. MTS’s N-drive alone is holding more than half a million files and folders containing more the 60,000 datasets. Furthermore, “IDC surveys find that the time spent searching for information averages 8.8 hours per week.” The integrated planning section typically invest significant amount of time searching for data relevant to a particular project, initiative or task, as well as hire outside consultants to provide that data. The goal of this Analytical Services Internship was to work with staff to develop a strategy for data discovery within the planning groups of the Air District, MTC and its transitional employees. The solution design is not to drastically change how things are done, but rather to improve and streamline the process with clear understanding of rules when it comes to data methodology to eliminate time, effort and cost spend on finding relevant data and greatly enhance the efficiency of the unit.

**Approach**

Interview with the Head of the Department and at least one main Power Users to understand the need.

Schedule 15-30 minutes one-on-one interviews with the main data users to understand how they use data.

Analyze responses from one-on-one interviews to draw conclusions on the initial questions and methods to gather information about datasets.

Investigate option for publishing survey: google forms, survey monkey, in-house webform.

Sent out initial survey as an in-house web form and as excel template providing incentive such as prize to complete with in shorter timeframe.

Schedule short one-on-one interviews with all users that didn’t respond to the survey.

Gather and document feedback on the Initial Data Survey.

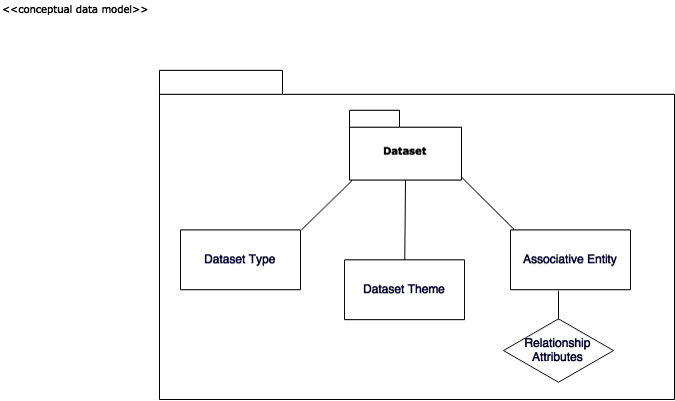
Analyze at least one source of data files that participants frequently referred to as location to evaluate whether or not comprehensive inventory of all data files/datasets is valuable:

Come up with Conceptual Framework Development

Document the Results

Present findings

**Conceptual Data Model Development Framework**



Dataset with Types and Themes and its Associative Entity (such as project or relationship to other datasets, projects, etc)

Data Types: Spatial, Non Spatial, Containing Spatial-Compatible field

Data Themes:

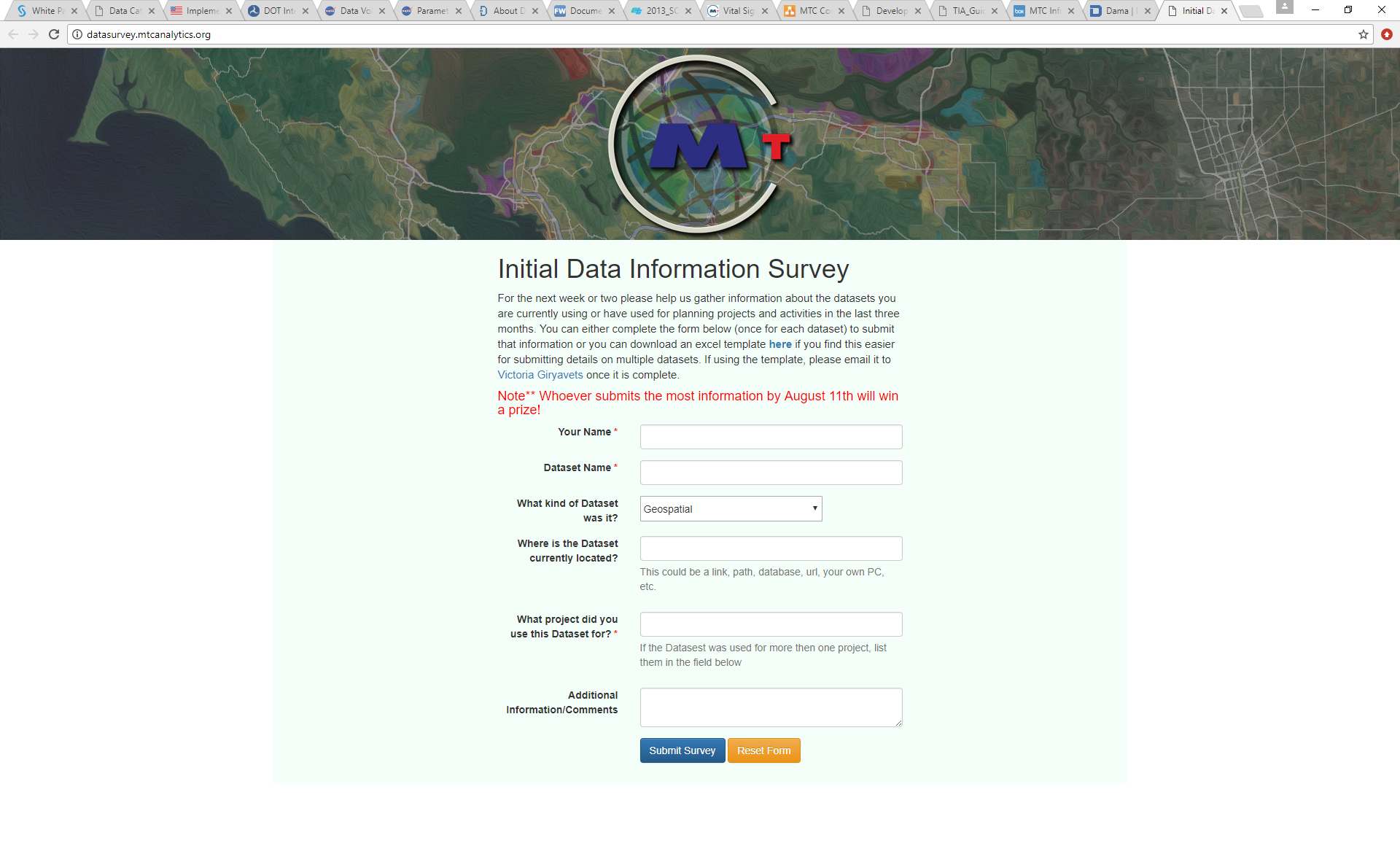
**Current:** Administrative, Demographic, Policy, Transportation, Environmental, Projects

**To consider:** Emergency Operation, Hazards, Structures, Government Units, Utilities, Addresses and Names, Transportation, Cadastral, Hydrography, Environmental, Land Use/Land Cover, Basemap, Elevation, Imagery, Geodetic Control

**Results**

**Initial Data Information Survey**

Webform <http://datasurvey.mtcanalytics.org/>



* Survey Participation results: 28% of people completed it or responded.(Michel Smith’s submission was past the Prize Deadline)

Datasets can be evaluated on a uniform set of criteria such as value, quality and manageability.

Value: 40%

Quality: 35%

Manageability: 25%

Approximately 30% of files actually should be inventoried and only 10-15% will be published. This trend might carry through any other source/location of datasets. N-Drive is an unstructured source of information thus makes it difficult to locate and use correct information.

|  |  |  |
| --- | --- | --- |
| **Time Frame** | **1 Year** | **4 Years** |
| Search for Data Cost | $ 410,010 | $ 1,640,038 |
| Reduced Cost (30 minutes less) | $ 293,530 | $ 1,174,118 |
| **Savings** | $ 116,480 | $ 465,920 |

**Comprehensive Data Inventory Cost and Benefit**

Evaluation

Techniques used

* In Person Interviews
* Web Survey
* Excel Template
* Mass email
* Personalized Emails
* Follow up strategy

What worked and what didn’t

**Conclusion**

With the information gathered via web survey, excel template, emails, in-person interviews and "drive-by"s it can be concluded that due to high volume of existing data and tasks it would be beneficial to have a mandatory Comprehensive Data Audit (not a survey where participation is voluntary) performed with the help of a Data Team. Aiding can be done via email or online Dataset Submission Form. It is highly advisable to develop a clear and comprehensive Data Organization Model with in the entire organizations that will help eliminate major cause of confusion when it comes to Data Quality Management, Data Development, Document and Content Management as well as Data Security and Data Operations Management. Flexible and Scalable Data Organization Model will eliminate skepticism and increase data integrity.

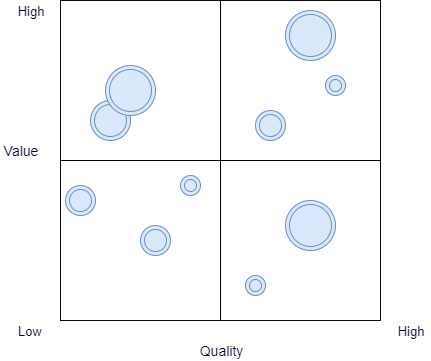
**Suggestions:**

Datasets can be evaluated on a uniform set of criteria such as value, quality and manageability.

Value: 40%

Quality: 35%

Manageability: 25%



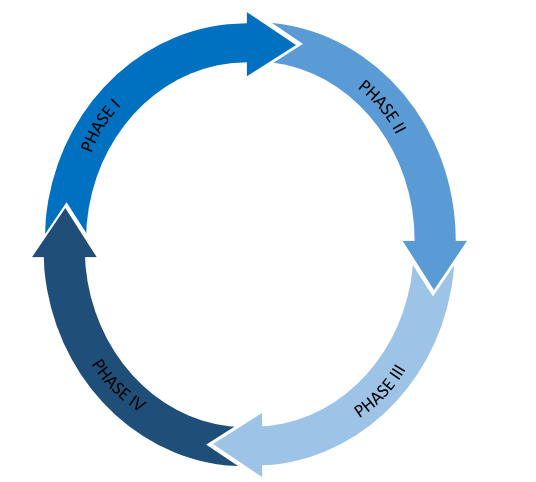
**Phases of Repeatable Data Organization Modeling Process**

Phase I: Preparation

Phase II: Collection

Phase III Evaluation and Organization

Phase IV Implementation and Publishing (share)



**Data Inventory Process Stages:**

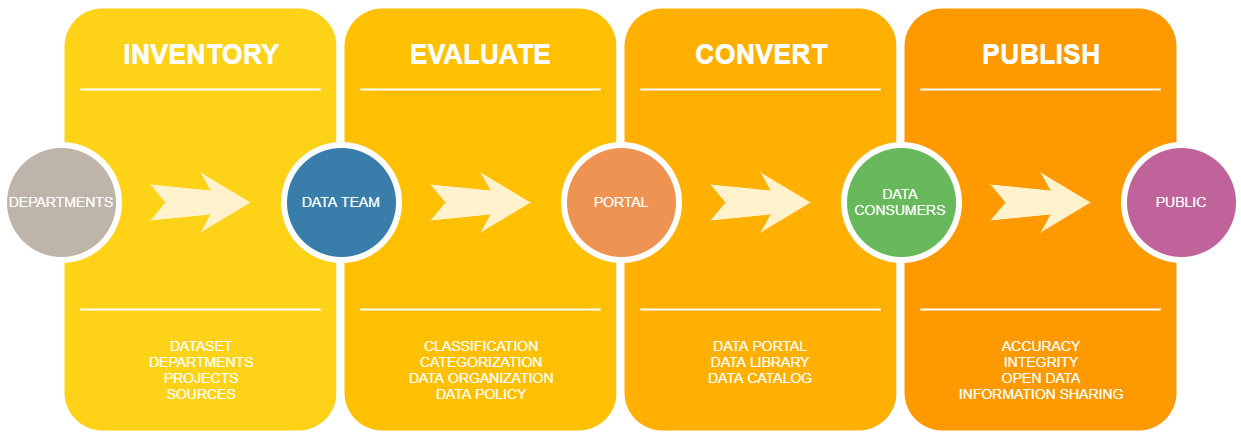
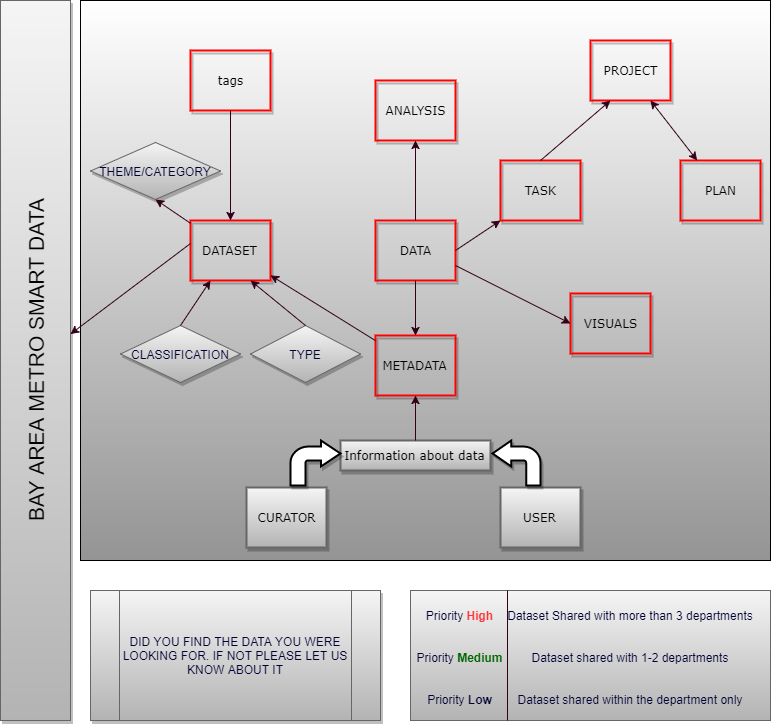


Figure 2: Victoria Giryavets (Analytical Services Intern)

**Smart Data Tracking, Initial Data Audit, Dataset Catalog, Data Library**

**Data Ontology** to bring everyone on the same level of understanding what is Data, Dataset, Data types, classifications and categories. Data Ontology is maintained by Data Administrator  
  
Examples of Ontology entries:  
  
**Data** - Representations of facts, concepts, or instructions in a formalized manner suitable for communication, interpretation, or processing by human or automated means. The fundamental components of information   
  
**Dataset** - a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

**Bay Area Metro Smart Data**



**Case Studies**

US Open Data Initiative

Massachusetts Department of Transportation

City of San Francisco

United States Department of Transportation

U.S. General Services Administration, Technology Transformation Service

National Aeronautics and Space Administration

Federal Geographic Data Committee

Valley Transportation Authority

State of New York Open Data

State of California Open Data

Chicago Regional Transportation Authority

Rudin Center for Transportation Policy and Management

LA Open Data